What works and what fails? Evidence from seven popular literacy 'catch-up' schemes for the transition to secondary school in England

Stephen Gorard, Nadia Siddiqui and Beng Huat See
Durham University
s.a.c.gorard@durham.ac.uk

Abstract

There are concerns that too many young people, from disadvantaged backgrounds, are moving into secondary education in the UK, and elsewhere, without the necessary literacy skills to make progress with the wider secondary school curriculum. A large number of interventions have been proposed to reduce this poverty gradient. This paper summarises the evidence from randomised controlled trials of seven popular interventions, giving a different comparative perspective to individual reports, and permitting more detail than a wider review. Of these, it shows that Switch-on Reading (Reading Recovery) and Accelerated Reader, for example, are currently the most promising. And that summer schools and the use of generic literacy software are the least successful and may even harm pupil progress. The way in which the evidence is assessed in this paper suggests a way forward for practitioners and policy-makers navigating the evidence in their areas of interest. There is also evidence that practitioners should be able to conduct robust evaluations of their own with only minimal support, which could lead to a revolution in school improvement. The combined results suggest that 'soft' evaluations may be worse than just a waste of time and money, and that theoretical explanations might appear satisfying to readers but are largely unnecessary when assessing 'what works' in education.

Introduction

In the UK, as in many other countries, there has been concern that some pupils are leaving primary education at age 10 or 11 without the basic skills needed to access the secondary school curriculum. Most importantly, perhaps, too many are considered not to have achieved the expected threshold level of literacy. Such children are not likely to catch up with, and are more likely to continue to fall further behind, their peers at school (West et al. 2005, Reyes et al 2000, Sainsbury et al. 1998, Galton et al. 1999) which can also lead to other issues such as anxiety, depression (Graham and Hill 2003) and disruptive classroom behaviour (Galton and et al. 2000). Underachievement at primary school is a strong predictor of pupils' academic performance at secondary level. If the underachievement persists during transition to secondary school, it is likely that the pupils will remain vulnerable to the risks of failure in future life chances and career opportunities. Catch-up literacy projects are educational interventions intended for such pupils struggling to reach what are officially deemed the age appropriate levels in reading (Gov.UK 2012). Such catch-up programmes are customised interventions, aimed at narrowing the reading achievement gap during and immediately after the transition stage.

Attainment in reading is an important concern of the Department for Education in England, and policy initiatives have been introduced in order to achieve the national targets for attainment in reading at Key Stage 2. One of the most important of these was the 2010 Pupil

Premium Funding policy. Under this scheme, schools are given extra funding in proportion to the number of pupil premium (PP) students they have. PP students are those officially defined as living in relative poverty, some children who had lived in care, and children of the armed forces. In general, PP students have lower levels of attainment than other students. Therefore, schools are expected to use this annual extra funding to improve the attainment of lower achievers and so reduce the PP attainment gap. As long as the funds are used for this purpose, schools have the freedom to build capacity and buy resources or teaching approaches that can support pupils who are at risk of under-achieving. What schools need, therefore, is reliable guidance on which approaches would be most suitable and effective for their own context.

This paper looks at a body of evidence on a range of literacy interventions which have been evaluated, some of which show promise of positive results. Some reading interventions appear to be effective, at least for some struggling readers, but some do not or have not been tested properly (See and Gorard 2014). Specifically, this paper reports summary evidence on seven of the most promising or widely used interventions that could be used either when pupils are getting ready to leave primary school (in Year 6), or during the holiday between primary and secondary school, or when they first arrive in secondary school (Year 7). The aims of all these interventions are to overcome the gaps in literacy learning and support pupils to join the mainstream group of learners.

All of these interventions have been implemented by school teachers and teaching assistants. The settings and context of implementation are English mainstream schools and classroom environments. The content and delivery approaches followed in these interventions are distinct from normal lessons because the intention is to support disadvantaged pupils through a catch-up programme, which might involve taking them out of lessons for small-group work, or providing extra resources. All of the evaluations reported here are independent of the sponsors and developers of the interventions, and conducted by the authors. Therefore, there is consistency in terms of following evaluation protocols and reporting the evidence as clearly as possible. Readers should note that, having been funded to evaluate the impact of a range of literacy catch-up schemes, we do not necessarily advocate any of them. It makes sense in many ways to deal with problems of poor literacy before Year 6 of primary education. However, where problems arise in Year 6 (and they do) the evidence from this paper can help direct practitioners and others towards the most promising at this stage.

The paper first summarises the general design and methods used for all studies, then introduces each intervention and summarises the prior evidence for it. Then we describe the specific methods used for each evaluation, before showing the headline results. The paper concludes by discussing the considerable implications for research, policy and practice. Its purpose, in presenting results from such a large number of studies, is to permit easier comparison between them and their outcomes. Therefore, each study cannot be presented in as much detail as it would otherwise. Among the issues covered are how to present and compare the results of different trials with overlapping aims, and whether schools can conduct robust evaluations of their own interventions.

Summary of methods used in the trials of seven popular interventions

To save space, this section presents the elements of research common to all seven studies. All of the studies used the same basic design, involving only two groups – the treatment group receiving the intervention and a control group receiving standard practice (Gorard 2013). In

all studies the cases were randomised to group by schools, classes or as individuals. Most of these (Switch-on, Accelerated Reader, Philosophy for Children, Fresh Start, literacy software, and Response to Intervention) were based on a waiting-list design in which the control group received the intervention once the trial was complete. The pupils were from year 6 (at the end of primary school) or the start of year 7 (in their new secondary schools), in state-funded schools from across England. All studies had both a pre- and post-intervention measure of literacy attainment (and sometimes for other subjects such as maths).

All analyses were based on intention to treat, meaning that the pupils were handled as being in the group they were randomised to, whatever happened subsequently, and pupils were followed up as far as possible even where they had moved schools. In addition, a sub-analysis of only those pupils eligible for free school meals (FSM) was conducted were possible. The analyses are based on the Hedges' g 'effect' size (the difference in means between the groups, divided by their pooled standard deviation). Where possible, the differences analysed are the gain scores from pre- to post-test, in order to cater for any slight imbalances in the initial groups, and to aid comparison between trials. If the pre- and post-test scores had a different metric then they were standardised as z scores before analysis. All scores are presented as rounded to eliminate decimal places and made comprehension easier. Further details can be found in the individual reports.

In order to help readers to assess the security of the findings, the design, methods and achieved samples for each study are rated between 0 and 4*, on a number of factors such as design, scale and attrition, as described in Gorard (2014) and summarised in Table 1 here. Each study is given a rating representing its lowest row description in Table 1 for any of the first five columns. All of the studies are randomised controlled trials – which is a good design for an impact study (row 1 of Table 1). All use either standard assessments such as Key Stage results, or standardised tests of attainment from GL Assessment which are independent of the intervention. Those studies with individual randomisation of pupils to groups are, all other things being equal, intrinsically superior to those where classes or schools are randomised. Otherwise, the larger the trial, and the lower the dropout, the more trustworthy the results are. Trials in which the groups were reasonably well-balanced at the outset are, all other things being equal, better than those where randomisation leads to imbalance. Other threats to the security of the findings are noted where relevant.

Table 1 – A 'sieve' to assist in the estimation of trustworthiness of descriptive work

Design	Scale	Dropout	Data quality	Threats	Rating
Strong design for RQ	Large number of cases (per comparison group)	Minimal attrition, no evidence of impact on findings	Standardised, pre-specified, independent	No evidence of diffusion, demand, or other threat	4★
Good design for RQ	Medium number of cases (per comparison group)	Some attrition (or initial imbalance)	Pre-specified, not standardised or not independent	Little evidence of diffusion, demand or other threat	3★
Weak design for RQ	Small number of cases (per comparison group)	Moderate attrition (or initial imbalance)	Not pre-specified but valid in context	Evidence of diffusion, demand or other threat	2★

Very weak design for RQ	Very small number of cases (per comparison group)	High attrition (or initial imbalance)	Issues of validity or appropriateness	Strong indication of diffusion, demand or other threat	1★
No consideration of design	A trivial scale of study, or N unclear	Attrition huge or not reported	Poor reliability, too many outcomes, weak measures	No consideration of threats to validity	0

In order to help readers further, each result is compared to the number of counterfactual cases that would need to be added to the smallest group in order for the apparent 'effect' size to disappear (Gorard and Gorard 2015). This involves creating a counterfactual score such as the mean score for the smallest group plus or minus the overall standard deviation for both groups. The SD would be added if the mean of the smaller group (in scale) were smaller than the mean of the larger group, and subtracted if the mean of the smaller group was the largest. This counterfactual score can then be repeatedly added to the smaller group of cases until the ES disappears (as it must eventually). The number of these imaginary counterfactual scores needed to make the ES disappear is the measure of sensitivity. A simpler way that does not require direct access to the datasets (only the means and standard deviations) would be to set NNTD as the absolute value of the 'effect' size multiplied by the number of cases in the smaller group in the comparison. The larger this NNTD is, the stronger the finding. Importantly, the NNTD can then be compared directly to the number of cases missing (through dropout, missing values, or non-response). Where the number of cases missing is trivial in comparison to the NTD, this shows that the result cannot be attributed to missing data alone.

Each trial also had an integrated process evaluation to monitor progress, observe testing, and assess fidelity to intervention. This aspect of evaluation does not necessarily reflect the final results of the impact evaluation itself. The information achieved through the process evaluation, from observation of training and operation, and interviews with staff, pupils and parent, helps to understand the intervention, and any barriers to its implementation. There is not enough space to describe these components in detail for each trial.

Evaluating seven popular literacy interventions

Switch-On (Reading Recovery)

Switch-on Reading is derived from a long-standing programme called Reading Recovery (RR). This is an intensive one-to-one intervention for the lowest performing children, widely used in the US, Australia, New Zealand and the UK. The What Works Clearinghouse (2013) found four small scale evaluations of RR that met minimal evidence standards (Baenen et al. 1997, Pinnell et al. 1988, 1994 and Schwartz 2005), and these had mixed results. More recently, Tanner et al. in (2011) and May et al. (2013) reported positive impacts from school-level evaluations. There has been little evaluation of RR in the UK, and less of Switch-on Reading itself. A weak evaluation with primary age children (Coles 2012) reported an effect size of +0.8. There was promise but no guarantee of success when our new trial was set up.

The model of Switch-on Reading being evaluated was provided for new Year 7 pupils in mainstream secondary school settings in Nottinghamshire. The intervention is a short-term individual reading programme for pupils who have not achieved Level 4 English at Key Stage 2 (KS2). The intervention was delivered over 10 weeks and consisted of regular 20 minute one-to-one reading sessions with Switch-on trained staff members. The intervention was conducted by staff including SENCOs, librarians, teachers, and mostly by teaching assistants. Each member of staff was trained, and looked after no more than four pupils. Each pupil was given a schedule in which to come out of one standard class per day for 20 minutes at a time for the Switch-on session. The schedule was arranged so that parts of different lessons were missed.

Switch-on Reading revolves around appropriately matched books that have been finely graded in bands and levels to provide small changes in challenge over time. These books had not been used with Year 7 pupils before and so one question was whether the pupils and staff found them suitable. Each Switch-on Reading session should have consisted of:

- Reading a familiar book (perhaps the first 100 words only)
- Discussion on the material, visuals, cover pages and blurbs of the books
- Invoking interest of students by involving them in talking about visual content
- Reading of the text and using the running record sheet for analysis of reading
- Feedback to the student
- Introduction to a new book

Therefore, each session incorporated revision of a familiar text, introducing new vocabulary, practicing phonics and also improving comprehension through questions and talking about the texts. In each session the student should read excerpts of text from four books.

At some point in the 20-minute reading session the member of staff recorded the reading assessment of the pupil on a sheet, and made an inventory of errors such as words missed, substituted with another, mispronounced, repeated, plus self-corrections and appeals for help. The form for recording these events and the rules for completion were standardised, and an integral part of the intervention. Part of the intervention also involved analysis of errors. The average number of errors was calculated, and determined which book set was followed next. After each book, the adult trainer praised the child when an effective reading strategy was observed, and prompted the student to use new strategies where behaviour had not been effective or advice had been ignored.

The evaluation of this intervention involved 19 primary schools in Nottinghamshire, ranging in size from around 600 pupils to over 1,500. FSM eligibility ranged from 6% to 30%. The schools identified 314 pupils eligible for reading support. Half were individually randomised to immediate support and the other half formed the control. This meant that each school was both a treatment and a control school. The Phase 1 intervention group of 157 pupils was involved in reading every day, aiming for at least 40 sessions in the minimum of 10 weeks. The Phase 2 group of a further 157 pupils continued with normal lessons and any interventions or programmes that were also available to Phase 1 pupils and that would have been used anyway in the absence of this evaluation. The pre-test was conducted at the outset, and the post-test was conducted before Phase 2 pupils received the intervention. One pupil did not register a pre-test score, and five pupils did not take the post-test. The New Group Reading Test (versions A and B) was used for the pre- and post-tests. The evaluators observed the post-tests in operation, because the staff and pupils were no longer blind as to who was in which group. Both the pre- and post-tests were conducted on-line to encourage

standard format and timing, to reduce the potential influence of staff, and to create instant results for the schools and evaluators. Overall, the two groups were reasonably well-balanced in terms of their background characteristics.

This evaluation has a reasonable cell size of individually randomised pupils, initial balance between groups, and minimal attrition. The evidence from it is listed here as 4* in terms of its robustness. Further details of the specific implementation of this intervention and the protocol are in Gorard et al. (2015a).

Accelerated Reader

Accelerated Reader (AR) is a web-based intervention produced by the Renaissance Learning Company, used by over 2,000 schools in the UK (Topping 2014). The What Works Clearinghouse (IES 2008) reported the results of a systematic review of studies on AR showing no visible effect on reading fluency, a mixed effect on comprehension and a possible positive effect on reading achievement. These results are based on two studies that fulfilled WWC standards for systematic reviews (Ross et al. 2004, Bullock 2005). Both of these studies are based on the STAR tests which are integrated in the AR programme, and cannot therefore be regarded as independent assessments (Krashen 2007). The rest of the research consists of simple snapshot surveys (Clark 2013), and weaker evaluations (Scott 1999), including some suggesting a substantial negative impact (Mathis 1996). More recently, a study in the US reported a small negative effect size (Nichols 2013), whereas Shannon et al. (2015) reported a positive impact. Overall, it is not clear from prior evidence that the implementation of AR at such a large scale in the UK can be justified on the basis of the existing evidence of effectiveness. A more robust trial was appropriate.

AR is a networked computer-based management programme intended to encourage pupils in independent book reading, and allow teachers to monitor pupils' reading levels and progress. Based on this information, the teacher's role is to support pupils in making an appropriate selection of books for reading, and to motivate them in achieving advanced reading levels. AR starts with a Standardised Test for Assessment of Reading (STAR), a 20-minute screening test that determines each pupil's 'optimal' level of reading comprehension. STAR can be conducted repeatedly and periodically to monitor pupil's progress. It is recommended on the Renaissance Learning Inc. website that teachers should conduct STAR three to five times in a year to follow pupil's gradual progress. The readability of a book is calculated taking into account the word count, average sentence length, average word length and word difficulty. There are over 160,000 books (fiction and non-fiction) available in the AR programme, allotted to bands on the basis of a readability formula.

Once an appropriate book selection has been made, pupils are given time in school to read independently. AR recommends teachers motivate pupil to read regularly, and finish reading the selected book promptly. AR suggests 30 to 60 minutes of independent reading time every day. There are around 156,000 quizzes in AR. These reading practice quizzes assess pupils' comprehension of the specific books they select to read. The format is generally multiple choice items that ask factual and inferential questions from the book. The quizzes are computer based and can be taken on laptop and tablets. Each pupil gets an individual login and password to have access to AR and complete the quiz. It is recommended that pupils take the AR quiz within 48 hours of finishing the book.

Four individual secondary schools proposed the intervention and evaluation of AR over a period of 20 weeks. The AR developers were not involved in any of these four proposals so it was decided that the schools should run the trial as a co-operative with advice from the authors. All schools were urban, mixed, secondary stage schools, with a high proportion of disadvantaged pupils. The schools selected their target groups on arrival in Year 7. A target group of 349 Year 7 pupils across the four secondary schools was identified on the basis of their prior KS2 scores (pupils at Level 4c and below in English). Of these, 323 were individually randomised to groups (180 to treatment, 163 to control). Three schools conducted individual pupil randomisation. One school randomised into treatment and control group by classes. The school that randomised classes had 119 pupils identified in the target group and they were already spread across different class groups (i.e. the usual classes for that school). The school claimed that it was not practically possible to individually randomise the pupils and conduct the intervention. The evaluators ran a separate group analysis for this school and found that the groups were well balanced in terms of KS2 scores before the intervention began.

Pupils in the waiting list continued the usual school activities. There was no chance of contamination because pupils in the control group had no access to the AR programme. No school dropped out from the trial. A total of 8 pupils (6 from the treatment group) did not provide a post-test score. The average KS2 scores of those who dropped out in treatment and control groups was about the same, and neither unusually high nor low, given the eligibility criteria. The findings are based on the post-test scores for the New Group Reading Test. There was no formal pre-test, and the initial balance of the groups is assessed in terms of their Key Stage 2 English scores.

This evaluation has a medium cell size of individually randomised pupils, initial balance between groups, and minimal attrition. The evidence from it is listed here as 4* in terms of its robustness. Further details appear in Siddiqui et al. (2015).

Philosophy for Children

Since it was developed in 1970 with the establishment of the Institute for the Advancement of Philosophy for Children (IAPC), Philosophy for Children (P4C) has become a worldwide educational approach, and something like it has been adopted by schools in 60 countries across the world, although the nature of the practice varies (Mercer et al. 1999). However, the evidence base so far has been weak, in terms of impact on attainment. An initial evaluation of the original scheme was conducted using a matched comparison design involving only 40 pupils from two schools (Lipman et al. 1980). Trickey and Topping (2004) conducted a review of existing studies suggesting consistent moderate effects on a range of outcome measures, and these seemed to endure (Topping and Trickey 2007). However, these studies were not very secure. Two more recent randomised trials found positive gains in terms of cognitive ability test scores (Colom et al. 2014, Fair et al. 2015), but the results for attainment were not assessed.

The main aim of our new evaluation was to determine the effect of the P4C programme on the Key Stage 2 scores of pupils who were in Year 5 when the schools were randomised and Year 6 by the end of the trial. P4C is not really presented as a catch-up intervention, but here the results are only considered for Year 6 reading attainment.

P4C aims to help pupils' to think logically, to voice their opinion, to use appropriate language in argumentation and to listen to the views and opinions of others. Pupils and teacher sit in a circle so everyone can see and hear one another. The teacher negotiates with pupils on guidelines on the conduct of sessions and the purpose is to set some basic rules of communication agreed by all pupils. The teacher then introduces the planned material they have chosen in order to provoke pupils' interest, puzzle them or prompt their sense of what is important. A minute of silence is followed by pupils in pairs sharing interesting issues and themes, or jotting down key words.

Children present their group's question so all can see and hear it. When all the questions are collected and recorded children are invited to clarify, link, appreciate or evaluate the questions prior to choosing one for discussion. When the listing of questions is complete, the next phase is to select a one as a dialogue starter. The selection is made by pupils using one of a range of voting methods. The discussion floor is then open for all to share their views.

Pupils participate in the discussion, building on other pupils' contributions, clarifying them, questioning them and stating their own opinions. Whether agreeing or disagreeing the rule is to justify opinions with reasons. Teachers will often prompt pupils to imagine alternatives and consequences, seek evidence, quantify with expressions like 'all', 'some' or 'most', offer examples and counter examples and question assumptions. The closing of the session involves last words from all pupils. Pupils might have the same opinion as in the beginning or it could have changed as a result to dialogue. Pupils are invited to sum up their views concisely and without contradiction from others.

To address the impact of this approach, we conducted a randomised controlled trial of P4C where the intervention was carried out for one complete academic year. The study involved 48 primary schools from London, Hull, Sheffield, Manchester, Hertfordshire, Staffordshire and Stoke-on-Trent in England. None had prior experience of using P4C. All schools had at least 25% of their pupils known to be eligible for free school meals. Of these 22 were randomised to the treatment group (772 pupils in year 5 at the outset and Year 6 by the end), and 26 to the control (757 pupils). The two groups were well-balanced in terms of sex, FSM-eligibility and SEN status. The intervention lasted just over a full academic year. Opt-out consent forms were sent by schools to parents to inform them of their child's involvement in the programme, outlining the purpose of the trial and the need to collect essential data while assuring them of the confidentiality of potentially sensitive data.

The individual results for KS2 reading, writing and maths were provided by the National Pupil Database (NPD) linked to unique pupil numbers (UPNs) supplied by all participating schools. The Department for Education matched the scores to the pupils for the evaluators. Because the KS1 pre-scores and KS2 post-scores were on different metrics both were converted to z-scores to assist comparability. Less than 10% of pupils with pre-test scores were missing a post-test score. The main outcome in assessing the impact was the English Key Stage 2 scores of pupils who were in Year 5 when the schools were randomised and Year 6 by the end of the trial. This evaluation has a large cell size, but randomised at school level, slight initial imbalance between groups, and just under 10% attrition. The evidence from it is listed here as 3* in terms of its robustness. Further details are in Gorard et al. (2016a).

Fresh Start

Fresh Start (FS) is a 'systematic synthetic approach' to reading, in which individual letters are sounded out within words, and these sounds then blended to form the pronunciation of the word, and so to 'read' it. When writing, the combination of sounds is said aloud and then converted to letters and written on the page. FS is produced by Read Write Inc., whose literacy programmes are cited by OFSTED (2010) as used by the 'best' performing schools. However, the prior evidence related to FS is weak, because relevant studies have often been small, non-randomised, with high dropout or poorly reported. A study by Brooks et al. (2003) intending to evaluate FS for use with low attaining pupils at Key Stage 3 (KS3) only managed to retain 30% of its initial 500 pupils, making any claims for the success of the intervention weak. One local authority in England adopted FS in all of its secondary schools for pupils not meeting or likely to meet expected levels of literacy (Lanes et al. 2005). The impact was never evaluated properly. Their 'evaluation' report shows that the approach was popular and considered effective by teaching staff, but the only evidence of impact came from before-and after-figures in one school with no true comparator. A later summary of reading interventions for KS3 included studies of FS, reporting effect sizes of +0.25 to +0.34 for reading comprehension (Brooks 2007). All of the samples were small, with one study having only 29 cases, and there was high dropout, with studies not clearly reporting the comparator groups, the allocation of cases, and whether the groups were equivalent at the outset. Overall, therefore, the direct evidence for Fresh Start is limited, and mostly from small-scale studies not randomising pupils to treatments. Given that the approach is widely used, a larger randomised controlled trial was appropriate.

FS is tailored to get pupils who have missed earlier opportunities to catch up with their peers so that they can participate in mainstream literacy activities without falling further behind. The complete FS resource pack includes module sets, assessment charts, magnetic sound cards, speed sound cards, sound charts and poster, lesson plans, pronunciation DVD for teacher, teacher training books and handbooks to support the delivery of FS. The modules are graded according to reading age, and in this trial FS was conducted three times a week for one hour each over 22 weeks. In addition to receiving the resource pack, teachers also took part in a two-day training workshop provided by the developers.

The programme begins with an initial assessment of pupil's phonics and word recognition, assessed individually by teachers. Pupils are put into four groups according to the initial scores to ensure homogeneity within the groups, which is believed to encourage progress. Depending on the individual pupils' progress, teachers may also provide additional 20-minute regular one-to-one sessions.

The ensuing phonic lessons involve the systematic teaching of 44 sounds in English, using a sound chart and Speed Sound Cards. Pupils practise blending the sounds through Sound Talk (sounding-out) by repeating the sounds after the teacher. This process is assisted using a number of learning aids such as picture cards, picture books, Fred puppet and talking fingers. Nonsense words are also used for pupils to practise independent blending of sounds. Pupils practise writing, although the letters are not mentioned by their names. There are 33 modules altogether and pupils start with different modules depending on their entry level. The modules are graded in six sets and each set consists of a pack of five booklets with different titles. Pupils are assessed after completion of each set to see if they are ready for the next module.

Three heads of school clusters in different regions of England (Harlow, Holderness and Telford) independently proposed conducting FS as an intervention. The funders (EEF) felt

that each cluster was too small for a feasible efficacy trial. So they suggested that each cluster run their own intervention but that they should be constrained to use the same evaluation design, and the results should aggregated by an independent light-touch evaluator. The independent evaluator would also train the school research leads, advise on design, oversee implementation and conduct a process evaluation from start to finish. This is therefore a school-led trial, and like AR in addition to the substantive results it provided evidence on whether schools and teachers can conduct robust research with advice.

At the start, the independent evaluators held a one-day workshop for the heads and research leads in each of the 10 schools. This covered the craft of conducting a randomised controlled trial. A key issue was how to randomise the eligible pupils into the two groups, making the allocation fair and without bias. A second workshop was conducted by the evaluators with the cluster and school leads before the post-test phase. This explained the conduct of the test process (the need for 'blinding' or at least observation to prevent bias), and how to calculate and interpret the results for each cluster. The schools reported that the workshops were very useful. Attendance was high. And the evaluators also found them informative about the kinds of challenges teachers faced when conducting research projects in their schools.

A target group of 433 eligible pupils from the fresh intake to Year 7 was identified by 10 secondary schools, based on them having KS2 scores at or below level 4c in English. On the basis of the pre-test when in year 7, all but 29 of the 423 pupils were reading at National Curriculum level 4c or below, and 237 were reading at level 3c or below. By the end of the intervention, a total of 8 pupils provided no post-test. Reasons for absence included that they had left the country, were long-term ill, suspended, or no longer attended the school and did not provide details of their new school. There were therefore 419 pupils in the final analyses, of which 215 were in the treatment and 204 in the control group. The main outcome measure was the New Group Reading Test, used as pre-test (version A) and post-test (B). As with all relevant evaluations described in this chapter, the headline findings are based on the 'overall reading score' provided by the software. This is used because our prior work has shown that there is floor effect created by the minimum achievable score when using the 'standardised age scores' (Gorard et al. 2015a).

This evaluation has a reasonable cell size, randomised at individual level, some initial imbalance between groups, and low attrition. The evidence from it is listed here as 3* in terms of its robustness. Further details are in Gorard et al. (2016b).

Literacy Software

It is now routine for most schools to use technology-based products such as software packages and websites in teaching and learning – for literacy and other core subject skills. However, solid evidence on the educational benefits of using generic literacy software in the classroom products is not clear. Rigorous intervention studies with suitable controls often find little or no positive impact from the use of technology-based instruction compared to standard or traditional practice. A number of studies and systematic reviews found that software packages had no positive impact on reading achievement (Borman et al. 2009, Rouse and Krueger 2004, Goolsbee and Guryan 2005, Dynarski et al. 2007).

This intervention involved a piece of poplar literacy software, widely used in schools to allow pupils to work at their own pace, provide regular progress updates, and permit the teacher to devote larger amounts of time to pupils most in need. We do not name here the software used

in the treatment (or its publisher). While regrettable, this is what was agreed at the outset, and it anyway makes little difference to the implications of this research. The publishers claimed that their reading software was 'award winning', and that if 11 year-olds worked on this program for one hour a day, spread over six weeks, the program will quickly improve their reading skills including single word reading, sentence reading and non-word reading. It also reportedly improves reading speed, reading fluency, vocabulary, comprehension and reading stamina. The software was designed to be used in conjunction with standard reading exercises, based on the National Literacy Strategy in England. It was aligned with National Curriculum standards, developed with guidance from some of the leading reading experts, and grounded in the most current research on literacy, using a carefully structured whole/part/whole approach to reading instruction. It has customised professional development ranging from CD-ROM and online courses to on-site workshops. A comprehensive Teacher's Guide with activities and lesson plans were included in the package. Ongoing technical support was agreed with the software publisher for the period of the trial. All treatment teachers received software training about how to use the software from consultants sent by the publisher. The training included a demonstration of the most effective ways of using the software.

The sample consisted of Year 7 pupils in state-maintained schools in Yorkshire agreeing to co-operate with the research and possessing a minimum level of technology access and support. Eight classes out of the total of 31 did not take part because at least one parent objected to their child taking part in the study. This left 23 classes containing 672 pupils at the outset. These were randomised to treatment (11 classes, 319 pupils) or control (12 classes, 346 pupils). No schools or classes dropped out. Four pupils moved to schools in another area before the pre-test, and a further three moved before the post-test. It was not possible to conduct an intention-to-treat analysis using these, since we could not follow the seven missing pupils. Nevertheless, their numbers are small and divided between both groups. A simple sensitivity analysis suggests that their inclusion could make no difference to the clear results of this trial.

The resulting 665 pupils were given a pre-test of their existing literacy levels in the first week and an equivalent post-test was given to both groups after ten weeks of teaching. The assessment was the Lucid Assessment System for Schools (LASS secondary). The intervention took place for 10 weeks, over a single term. The control group remained in routine teaching practice using a more traditional paper and teacher based format, with no specified ICT component. The treatment group used the computer software for a designated time on three to four days each week. Headphones were supplied for every pupil to counter distraction, thereby maximising the pupils' attention.

The software, the treatment schedule and the training all encouraged teachers to help pupils complete all of the learning activities provided by the software, over the ten weeks of implementation. The software itself automatically logged the records of each activity completed by each pupil and class. Most pupils in all classes completed the bulk of the activities. One class had some technical difficulties with their computer system early in the term.

This evaluation has a large cell size, and low attrition, but is randomised at class level, and has some initial imbalance between the groups. The evidence from it is listed here as 3* in terms of its robustness. Further details are in Khan and Gorard (2012).

Response to Intervention

RTI is a personalised and targeted intervention developed in the United States as part of an inclusion policy to provide a differentiated programme of instruction for children with learning disabilities within regular school settings. The theoretical and empirical framework of the approach was based on work by Clay (1991) and Fountas and Pinnell (1996). According to Clay children learn literacy skills by developing an inner control of strategies for processing text. If a piece of text is too difficult, the child cannot develop this control. So any text used should be pitched at the right level. With effective and explicit teaching, the teacher can help the child build a strategy to enable them to process the text. Based on their work on Reading Recovery, Fountas and Pinnell (2006) developed an approach called Guided Reading using books matched to children's abilities employing differentiated instruction in small groups, gradually building up the child's inner control. This was the basis for the differentiated levels or tiers that forms the basis of the RTI approach.

Early evidence suggested that this approach was effective with pupils of transition age (Vaughn and Fletcher 2012), with a positive effect for pupils with severe reading difficulties, although the gains were not big enough to close the gap with typically performing pupils (Leroux et al. 2011). In a quasi-experimental study, Graves et al. (2011) suggested that RTI was particularly effective for pupils from disadvantaged backgrounds with difficulties in oral fluency. Faggella-Luby and Wardwell (2011) also reported positive results for the small group intervention for older children but not for the younger ones. There is promise but the approach has not previously been tested at scale and in the UK.

RTI is targeted at the specific needs of students in the form of a whole class approach as preventive teaching (Tier 1), followed by small group remediation (Tier 2) for those who needed more attention and one-to-one tutoring for those who do not respond to the small group instruction (Tier 3). The RTI programme in this study was designed by the Centre for the Use of Research and Evidence in Education (CUREE) who developed the specialist tools and resources, and delivered the training. Training was conducted prior to the implementation of the programme and after schools had been randomised. The training was a 3-day event which included an introduction to the concept of RTI, and the range of tools and protocols. Teachers were shown how to use these in screening pupils for eligibility and assessing their needs, and how to select appropriate research-based approaches. In addition, treatment teachers also received on-going support provided by another organisation known as AfA3As (Achievement for All 3As) through in-school coaching using their Achievement Coaches as part of the AfA programme.

Initially 91 schools were approached through the AfA3As network of schools. Of these 85 indicated interest, but 24 subsequently declined to participate when they realised what was expected of them, leaving 61 schools. After schools were recruited, all Year 6 pupils in the 61 schools (pupil N=2,352) took the New Group Reading (NGRT) pre-test (a standardised test of literacy). Schools were then randomised, with 30 allocated to receive treatment and 31 to a waiting-list control.

All schools (control and treatment) were then meant to identify and report to evaluators their eligible pupils (those who were at risk of not achieving Level 4 and likely to benefit from the intervention) using a combination of teacher's judgement about which child or group of children would benefit from the treatment and the NGRT data. In general, six to eight vulnerable target pupils were to be identified for each Year 6 class. Although not ideal, this

sequence relative to school randomisation and testing was adopted at the request of the developers who wanted to use the pre-test results to identify eligible pupils, and with the permission of the funders, but against the advice of the evaluators. What happened in practice was that the list of eligible pupils provided by schools was neither complete nor clear. There were 79 pupils from the control schools and 37 from the treatment schools whose eligibility was unknown. This was partly a consequence of the sequence and partly due to the lack of direct communication between the evaluators and the schools – both insisted upon by CUREE. Therefore, the analysis presented later is based on all pupils known to evaluators to be at or below Level 4c from the outset on the basis of the pre-test.

After randomisation, 11 schools (three treatment and eight control schools) dropped out, reportedly due to organisational issues as a result of changes in leadership. This is very high, and at a scale not encountered by the evaluators before. In addition, one other control school conducted the post-test on the wrong year group of pupils. So valid data from only 49 schools was analysed (27 treatment and 22 control). Overall attrition was in excess of 25% meaning that the results of the trial must be treated as indicative only. The findings are based on the gain scores from pre- to post-test for the New Group Reading Test.

This evaluation has a reasonable cell size, but randomised at school level, initial balance between groups, but then high attrition. The attrition means that the evidence is listed here as only 2* in terms of its robustness. Further details are See et al. (2015).

Summer School 2013

In 2012 and 2013, Future Foundations ran a summer school in disadvantaged areas, based on the style used in the US by Building Educated Leaders for Life (BELL), a style reported as one of the few forms of summer schools with reasonable evidence of success (Terzian and Moore 2009). There have been several evaluations of the BELL summer schools in the US (BELL 2001, 2002, 2003). Unfortunately, their own 'evaluations' are often unclear, and reinforce the importance of independent evaluation where concern is more about finding the impact than in what that impact is (see Harvard Family Research Project 2006). In their study, gains are reported but no effect sizes were published. The gains were lower for low-income children and those in the age range relevant to school transition. A similar study by researchers with a potential conflict of interest looked at BELL summer schools in two US cities (Chaplin and Capizzano 2006). The overall 'effect' size for reading was calculated and found negligible. More importantly, 46% of those randomised dropped out or refused to continue with the study, and the results are available for only 44% of the initial randomised students. Overall therefore, despite some claims to the contrary, there is no strong evidence that the BELL approach would work in England with disadvantaged pupils preparing for secondary school. A synthesis of 93 evaluations of summer schools more generally suggested that they can be effective, especially with parental involvement, but perhaps with more promise in maths than literacy (Cooper et al. 2000). Schacter and Jo (2005) and Matsudaira (2008) found positive effects of summer school programmes for literacy gains of economically disadvantaged first grade children in the US.

In both 2012 and 2013, funded Future Foundations ran a summer school in disadvantaged areas, based on the BELL. The programme was intended to target pupils from disadvantaged backgrounds, who were underperforming at their expected or potential levels, and likely to benefit from participation in the programme.

In 2012 the school took 160 pupils who had completed years 5 and 6. One of the reasons for this pilot programme was that there was little robust evidence on the efficacy of the BELL approach in the UK. The pilot was therefore designed to test the feasibility of organising a summer school in a relatively deprived area. In particular, it sought to assess: whether there was demand for the programme, whether families would support and sustain the programme, and whether professional staff would be willing to work during their summer holidays. The report suggested that the approach was feasible, in a suburb of London (Siddiqui et al. 2013), and so a larger trial was set up in more varied locations to establish impact on attainment. In 2013, the plan was to take 1,000 pupils at the end of the years 5 and 6 in three separate settings.

Pupils attending the four-week programme followed a specially designed curriculum involving regular literacy and numeracy lessons taught by trained teachers. Lessons were supported by mentors and peer-mentors and generally conducted in small teaching groups. Each afternoon, students participated in a variety of sports and enrichment activities. The programme took place across three sites in London and the South East: Brighton, Enfield and Islington in the summer of 2013. It was targeted primarily at pupils in Years 5 and 6 who were eligible for free school meals (considered to be disadvantaged by their school) and/or who were not expected to achieve Level 4 in English or maths at the end of Key Stage 2.

In total, only 435 pupils (and their parents) volunteered to take part in the second study, suggesting perhaps that the treatment was not that attractive. The randomisation resulted in the allocation of 239 pupils to the treatment group to attend the summer schools programme, and 196 pupils to the control which means they were simply followed for the post-test. In the final analysis the sample retained had only 303 pupils – with 75 treatment and 30 control pupils not included in the final analysis. The former were largely those allocated to treatment but then not turning up to the summer school. The reasons given were that parents had work or holiday arrangements that clashed, or pupils were ill, changed their minds, or did not want to attend as their friend(s) had not been selected. The missing control pupils were largely those who moved away or whose subsequent secondary school would not conduct the test.

The 435 cases were randomised (239 in the treatment group and 196 in the control). The prior attainment scores consisted of KS2 fine point scores from summer term 2013, and the post-test was Progress in English administered in groups in the secondary schools attended by both groups of pupils in autumn 2013. Primary schools were generally co-operative in conducting the tests. It was harder to get agreement from secondary schools to test the original Year 6 pupils after they had begun Year 7. A great deal of effort was put in to reduce demoralisation and consequent dropout. This involved not revealing the groups until after the randomisation, use of a refundable deposit for registrants, and neutral administration of the post-test. Originally, the evaluation was intended to include an analysis of dosage. However, the attendance figures provided by the developers were not complete enough to conduct this.

This evaluation has a reasonable cell size, randomised at individual level, and initial balance between groups, but high attrition. The evidence is listed as 2* in terms of its robustness. Further details are in Gorard et al. (2015b). This study was preceded by a smaller study of the same intervention in 2012, with the same design but rated 1* (Siddiqui et al. 2014), and its results are also described later.

Which interventions were the most effective?

There is only sufficient space here to present the headline outcomes for each of the trials described so far. Further analyses for other outcomes and sub-groups can be found in the individual report for each trial.

Switch-on (Reading Recovery)

Switch-on is considered first here, because its evaluation is the most robust. Overall, the effect size of the intervention was +0.24, suggesting a noticeable positive impact (Table 2). Both randomised groups had very similar scores at the outset (NGRTA), which suggests that the randomisation was effective and so the test of the intervention was fair in that respect. The result based only on pupils eligible for free school meals (an official indicator of poverty) was an 'effect' size of +0.36.

Table 2 - Difference in gain scores for Switch-on Reading Programme

Treatment group	N	NGRTA	NGRTB	Gain	Standard deviation	'Effect' size
Switch-on	155	77	81	4	8	+0.24
Control	153	76	79	3	7	-

The number of counterfactual scores needed to disturb the finding would be 37, and there are only five cases missing post-scores. The headline finding of this study is therefore that the intervention is effective overall, and especially for disadvantaged pupils.

The two-day training event for staff from all schools was professional and successful, followed by on-going support and school visits from the developer. The evaluators observed most pupils having made considerable progress both in terms of the band of books and their reported reading age. Most pupils reported enjoying the sessions, and staff were generally positive about the programme. There were some concerns that the books were not suitable for secondary school, and these were altered.

Accelerated Reader

In terms of their prior KS2 English points, the randomisation was successful in creating balanced groups at the outset, which means that the analysis produces equivalent results whether it uses gain scores or post-test only (Table 3). Considered in terms of the NGRT reading scores the treatment group is ahead of the control by about one quarter of a standard deviation at the end of the trial, suggesting that AR has had a modest impact on the treatment group. An analysis using only those pupils listed as eligible for free school meals produced an 'effect' size of +0.38.

Table 3 – Prior KS2 points in English and NGRT outcomes, by treatment group

Group	N	KS2 points	Standard deviation	NGRT	Standard deviation	'Effect' size
Accelerated Reader	174	27	4	327	51	+0.24
Control	161	27	4	315	47	-

The number of counterfactual scores needed to disturb the finding would be 39, and there are only eight cases missing post-scores. The headline finding of this study is therefore that the intervention is effective overall, and especially for disadvantaged pupils.

Overall, schools were observed to be implementing the intervention faithfully. Two of the schools purchased tablets to make the AR quiz time a more fun activity for pupil. In terms of AR implementation during the transition from primary to secondary school one of the school leaders reported that it was not the appropriate timing for some of their pupils to be introduced to AR. Pupils coming from primary schools needed support from teachers to adjust to the new format of secondary schooling. Facing them with AR, in addition, could be a challenge at this beginning stage of a secondary school experience.

Philosophy for Children

At the outset the treatment and control groups were slightly unbalanced, with the control group having better KS1 scores in reading (Table 4). By the end the treatment group had narrowed this gap, leading to an estimated 'effect' size of +0.12. Based only on FSM-eligible pupils the 'effect' size was +0.29. There is evidence here that P4C might have a positive impact on pupil attainment at KS2.

Table 4 - KS1 to KS2 reading progress, by group

	N	Mean KS1 points z-score	SD	Mean KS2 fine points z-score	SD	Gain z-score	SD	'Effect' size
Treatment	772	-0.08	1.01	-0.02	1.01	+0.06	0.88	+0.12
Control	757	+0.08	0.98	+0.02	0.99	-0.05	0.91	-

The number of counterfactual scores needed to disturb the finding would be 91, and there are no cases missing for KS2 post-scores and no school dropped out of the intervention or the evaluation. The headline finding of this study is therefore that the intervention is probably effective overall, and more so for disadvantaged pupils. The lowest attaining half of the pupils did not improve their scores more than the higher attaining half, across both groups combined. Therefore, the result cannot be due to regression to the mean.

The intervention was appealing to many schools as a way of raising and debating pupil-school discipline problems in an enquiry group. The school leads reported that they discussed the concepts of bullying, racism, lying and cheating, equality and fairness which are core issues of school discipline and ethos. P4C was reported by the teachers to be very helpful in pupils thinking critically about these issues, raising questions, reflecting on their experiences and coming to fair conclusions. Some of the examples of questions discussed in P4C observed sessions created by pupils themselves from the given stimuli were as follows:

- Is it acceptable for people to wear their religious symbols at work places?
- Are people's physical looks more important than their actions?
- Can you and should you stop free thought?

There were some clear challenges to the delivery and implementation of P4C – such as the difficulty of embedding P4C in the fully-packed timetable and with targets for literacy and

numeracy from the National Curriculum, and a danger that the approach may be open to the influence of teachers' biases, beliefs and ideologies.

Fresh Start

The control group was ahead in terms of reading from the outset, making the gain scores a fairer test of impact here (Table 5). The intervention showed a small positive impact on reading comprehension (+0.24). The same 'effect' size occurred when only FSM-eligible pupils are considered.

Table 5 - Gain scores for Fresh Start reading

	N	NGRTA pre-test	Standard deviation	NGRTB post-test	Standard deviation	Gain score	Standard deviation	'Effect' size
Intervention	215	252	65	280	60	28	48	+0.24
Control	204	274	58	291	53	17	42	-

Because of the initial imbalance between groups, we have to treat the results as slightly more tentative than if the randomisation had led to more equal average scores. However, the lowest scoring pupils in the treatment group had a slightly lower gain score than the lowest scoring pupils in the control group. This suggests that overall the result was unlikely to be due to regression towards the mean.

The number of counterfactual scores needed to disturb the finding would be 49, and there are only 10 cases missing post-scores. The headline finding of this study is therefore that the intervention shows promise of being effective, but not especially so for reducing the poverty gradient in literacy.

The teachers attended a two-day training workshop given by experienced and professional trainers. The classroom management strategies are inspired by reception years and primary school teaching. FS teachers are expected to use body language, praises and dramatisation to get pupils' attention. As such, several secondary school teachers found these strategies and management styles difficult to adopt. Similarly, one school leader reported that parents expressed initial concern that the intervention was too low level for secondary school pupils. There was also some initial resistance from some pupils who felt that the activities were patronising. There was also resistance from some of the more experienced teachers. In fact a head of English in one school refused to take part, and found a substitute. However, classroom observations suggested that the FS teaching strategy was quite effective in keeping pupils engaged. The pupils received a lot of support and individual attention which they would not otherwise have had. The attendance records showed that pupils were attending the sessions regularly. Many pupils reported that they preferred coming for these sessions rather than regular classes.

Literacy software

At the outset of the trial, the pre-test scores show that both groups had similar literacy levels, with the treatment group slightly superior (Table 6). However, by the end of the trial the control group had caught up and overtaken them, with an overall 'effect' size of -0.29. This suggests that the intervention was harmful to pupils' reading compared to normal teaching.

Table 6 – Pre- and post-test scores for both groups

	N	Pre-test mean	SD	Post-test mean	SD	Gain score	SD	'Effect' size
Treatment	319	823	68	863	88	40	79	-0.29
Control	346	817	72	886	78	69	131	

The number of counterfactual scores needed to disturb the finding would be 93, and there are only 16 cases missing post-scores. The headline finding of this study is therefore that the intervention shows no promise of being effective, and is very likely to be harmful. No analysis in terms of FSM-eligibility is possible.

Intriguingly, the in-depth data collected routinely as part of the trial suggested a high level of satisfaction with the treatment. The technology-based instruction reportedly provided teaching groups with a range of information, links and activities in an accessible and entertaining way. The teachers involved in the treatment said that the software had an encouraging focus on language for early Key Stage 3 and that the activities were stimulating for pupils and teachers alike. They believed that it offered a reliable way to help pupils improve their reading skills. The pupils were satisfied with the technology-based reading materials, and were observed getting heavily involved in the activities. When asked, all teachers indicated that they would use the same or similar software in the future, and almost all of them said that they would recommend it to other teachers.

Response to Intervention

In terms of the reading test, RTI appears to have had a modest positive impact (+0.29). The two groups were reasonably well-balanced at the outset, and so a post-test only analysis produces a similar result (Table 7). An analysis using only those pupils listed as eligible for free school meals produced an 'effect' size of +0.48.

Table 7 – Pre, post and gain scores for RTI trial

	N	NGRTA pre-test	Standard deviation	NGRTB post-test	Standard deviation	Gain score	Standard deviation	'Effect' size
Intervention	171			287	51	22	50	+0.29
Control	180			270	60	16	42	-

However, all of the results have to be taken as indicative only, because of the level of school dropout after allocation (25%), the number of schools in the control group which did not carry out post-testing, and the inconsistency with which pupils who were eligible for the intervention were identified (see above).

The number of counterfactual scores needed to disturb the finding would be 50, and there are 166 cases missing post-scores because 11 schools dropped out of the intervention and as well as from the evaluation. The headline finding of this study is therefore that the intervention shows only weak evidence of being effective.

There were wide variations in the intensity and frequency with which schools implemented RTI, and in some schools the time allocated was too short for proper monitoring, tracking and adjustments to intensity. This was largely due to the timing of the programme, being introduced in the last few weeks of the final term. Accounts from teachers, pupils and Achievement Coaches suggested that RTI has beneficial effects on pupils' literacy as a catch-up literacy programme. One school claimed that the data they collected showed that their

pupils had made an equivalent of five months' progress in comprehension and spelling in the four weeks, and in some cases as much as a year's progress. The teacher reported how two pupils improved their reading fluency from reading 200 words in three minutes to nearly half of that time.

Summer School 2013

The two treatment groups were reasonably well balanced in terms of the available prior performance from the outset. Therefore, whether considered as post-test only or gain scores, the 'effect' size is around +0.17 (Table 8). The results based on only FSM-eligible pupils are the same.

Table 8 - Gain score English, all achieved scores

	N	KS2 reading	Standard deviation	PiE score	Standard deviation	Gain score	Standard deviation	'Effect' size
Treatment	169	23.66	4.86	39.06	16.08	3.71	13.24	+0.17
Control	144	23.42	5.76	36.36	15.41	1.46	12.79	-

The number of counterfactual scores needed to disturb the finding would be 25, and there are 122 cases missing post-scores. The headline finding of this study is therefore that the intervention shows little robust evidence of being effective.

Staff training was conducted for all teachers, mentors and peer-mentors which was well conducted and structured, with many examples of activities which teachers could use. Teachers were generally motivated and eager to try something new. The planned literacy and numeracy sessions were closely followed on all sites. The implementation of literacy sessions was very close to the developed lesson plans. However, numeracy sessions were observed to change with time into individualised tutoring sessions. One of the reasons was perhaps a wide range of abilities in some class groups, and teachers with the help of mentors broke numeracy classes into further small groups or one to one sessions. Several lessons observed by the evaluators were poorly taught - especially for maths. Basic pedagogical and factual errors were observed, and in one case pupil written responses were marked incorrectly. In literacy especially, more sessions were seen to be fun and enjoyable for all. Despite the low pupil ratios, class control was sometimes poor. Pupils were generally enthusiastic about attending the summer schools. Most of them also reported that it was the afternoon activities they enjoyed more than the teaching sessions.

Summer School 2012

Table 9 presents the summary results for all eventual Year 7 pupils for whom there are pre- and post-intervention results. Surprisingly, the KS2 score in reading is higher on average for those pupils attending the summer school than the control group. This initial difference is not large, but it does raise the question of to what extent the summer school catered for the lowest attaining and most disadvantaged pupils. Both groups improved their average scores slightly over the summer, and the effect size of -0.02 suggests very little difference between the groups but certainly no advantage for those attending the summer school.

Table 9 – Estimated impact of Summer School Programme on Year 7 Reading

Treatment group	N	KS2 raw score	September raw score	Gain	Standard deviation	'Effect' size

Summer School	34	33	35	+1.5	5.6	-0.02
Comparison	53	32	34	+1.7	9.7	-

The situation for Years 6 pupils is similar but worse (Table 10). As with Year 7, there is no clear evidence from this data that the summer school catered for an especially disadvantaged set of pupils from these schools. The number of cases in each year is too small to present an analysis using only FSM-eligible pupils.

Table 10 – Estimated impact of Summer School Programme on Year 6 Reading

Treatment group	N	August score	September score	Gain	Standard deviation	'Effect' size
Summer School	22	22	20	-1.9	3.7	-0.14
Comparison	33	22	20	-1.4	3.6	-

Given the small size of the summer school group for whom scores were provided, and the scale of missing data, this is not definitive evidence of a harmful impact from attending the summer school, but it cannot be construed as evidence of any beneficial impact for either Year 6 or Year 7 pupils.

The number of counterfactual scores needed to disturb the finding would be 1 (Year 7) and 3 (Year 6), and there are 55 cases missing post-scores. The headline finding of this study is therefore that the intervention shows no evidence of being effective.

Conclusions

Table 11 provides a simple summary of the results so far, adding also the cost of each intervention. The costs can only be estimated from information provided by the developers, and are based on direct costs such as resources, equipment, licences and training. In some cases there are start-up costs that would reduce as a proportion over time. Some figures are dependent upon the number of pupils and the costs would reduce with more pupils. Some involve the use of staff time, such as the involvement of teaching assistants. Where possible, these staff costs are not included. However, the actual cost for an intervention like Switch-on would be considerably lower where it used books and staff members already at the school.

Table 11 – Summary of findings

	Effect size	Effect size FSM-only	Quality of evidence	NNTD-attrition	Cost per pupil
Switch-on	+0.24	+0.36	4*	32	£627
Accelerated Reader	+0.24	+0.38	3*	31	£9
P4C	+0.12	+0.29	3*	91	£16
Fresh Start	+0.24	+0.24	3*	39	£116
Literacy software	-0.29	-	3*	77	£10
RTI	+0.29	+0.48	2*	0	£175
Summer school 2013	+0.17	+0.17	2*	0	£1,370
Summer school 2012 Year 7	-0.02	-	1*	0	£1,400

Summer school 2012 Year 6	-0.14	-	1*	0	£1,400

Several conclusions are immediately obvious. The security of any finding does not depend only on the research design – or put more simply, a randomised control trial is not a 'magic bullet', and its results are not necessarily 'gold standard'. Here the studies with similar designs range from excellent (4*) to weak (1*), largely due to variation in attrition, and factors such as errors in identifying eligible pupils that were beyond the control of the researchers. Nor, assessed correctly, is the security of a finding linked to its 'effect' size. Robust studies can have negative or neutral findings, in the same way that weaker studies can have larger 'effect' sizes.

On the basis of these findings, a school looking to assist pupils with literacy at the transition period, and reduce the attainment gap between disadvantaged pupils and their peers, would be advised to select Switch-on Reading, or perhaps Accelerated Reader. This is justified by the impact, cost and security of the initial findings. However, each of these findings should still be replicated by other researchers where possible.

If improving the reasoning of children is appealing to a school for other reasons, then Philosophy for Children also looks a good 'bet' for reducing the poverty gradient in reading somewhat. Fresh Start is promising but there is no indication yet that it will reduce the poverty gradient in reading.

It is clear that simply using generic commercial software to teach literacy does not work, and this should be avoided.

Response to Intervention holds enough promise of impact and of being effective for poorer children for it to be assessed again, with individual randomisation and a different developer. Until then, the evidence here is not sufficient to suggest that it should be preferred to any of the alternatives above.

The summer schools did a lot more than teach literacy, but assessed purely in terms of impact for reading they were very expensive and hold little or no promise of reducing the poverty gradient.

More generally, the most successful interventions were based on individual or small-group sessions. It would be best if these were conducted as part of general literacy classes (where other pupils could have more advanced interventions or use the library), rather than the target pupils missing other lessons to attend the intervention session. The pupils would not then miss important lessons such as maths, or lessons they clearly enjoyed such as PE, and they would not face the potential embarrassment of being called out of regular classes.

Can schools conduct their own trials?

Two of the trials (AR and Fresh Start) were set up as aggregated trials where a number of schools with similar interests agreed to run the interventions by themselves (not using the developers), and to try and evaluate the impact for themselves. We were assigned as independent overseeing evaluators for both trials, to advise the school leads on the process of conducting research, randomisation and testing, and chiefly to aggregate the eventual results from all schools. The direct cost to the schools was zero, and the light touch independent

oversight cost the funders considerably less than a full trial would. As can be seen from Table 11, the resulting evaluations while only medium in scale were at least as good as those involving the developers, and in which the schools had no part to play in the impact evaluation.

The main advantages of schools running their own trials include their ability to monitor pupil attendance and progress, automatic access to the personal and possibly sensitive data that schools record, no school dropout, no communication problems with other parties, lack of vested interest, easy permission to innovate, and building the capacity of practitioners in reading and critiquing research claims. If conducting such research was seen as a part of schools' functions then the overall cost of research could go down. It may even be possible to create some kind of nationwide ongoing trial with all willing schools contributing to an on-line database, which could adjust its synthesis of evidence with each new (small) study, much in the way suggested for medicine by Goldacre (2012).

On the other hand, school leaders did not always appreciate the importance of some aspects of the evaluation. For example, when pressed they were happy to support the evaluators who were trying to locate and test missing pupils. But they did not do this on their own initiative, and had no real concept of the dangers from attrition (despite discussion of this in their training days). As another example, some may not have fully appreciated the importance of randomisation, and felt that re-randomisation was a solution to less than ideal allocation to groups. Although diffusion from treatment to control was not an issue in these trials it could in other interventions, and again it is not clear that school leaders fully understand the problems this may cause. It was observed that most staff involved became advocates for their programmes increasingly during the trial, and schools had already made arrangements to continue with and expand its use for future years. They did not all have the mental equipoise needed to conduct a fair test.

Does the precise intervention protocol matter?

Intriguingly, three of the most robustly evaluated interventions yielded 'effect' sizes of +0.24. This leads to the question of whether the precise protocols specified by the developers of these three interventions are actually the 'active ingredients' of any success. All of these interventions are based on underlying approaches to learning such as small group teaching. Perhaps many coherent approaches and structured methods with time and resources equivalent to these, undertaken by volunteer schools (as all must be to have taken part in the trial), would be similarly effective for pupils at risk of failure. Maybe the exact nature of the intervention is irrelevant. If so, this would give practitioners more freedom to select from among these successful approaches the one(s) that best fit their context.

Does theory matter?

Each of the interventions in this paper has a plausible theoretical explanation as to why it should work. The summer school, for example, provided additional direct tuition time from selected successful teachers, during the period over the summer when there is traditionally a learning 'loss'. Yet there is no solid evidence that this raised the attainment of the pupils who participated. The literacy software was admired by all those involved. Pupils liked it because they could work at their own pace. Teachers liked it because it freed them, and allowed them the flexibility to work on a one-to-one basis for an extended period with pupils who needed it. School leaders and parents liked it because the regular assessments provided evidence of

progress. And yet, the pupils who did not use the software still made more progress. As noted in the context of a much larger review of evidence, it seems as though having a plausible theoretical explanation does not matter that much when considering what works (Gorard et al. 2011). If an intervention does not work, no amount of theory can salvage it in that form, but if it does work it does not really matter at that stage whether exactly how it works is understood.

Does 'qualitative' evaluation work?

As explained earlier, all of these trials involved a process evaluation consisting of observations of the interventions in practice, and interviews with stakeholders, such as school leaders, staff, pupils, parents and developers. The advantage of process evaluation is to assist explanation of the results, rather than determining the results of impact evaluation. The process evaluation cannot suggest if the intervention will work or not. It was remarkable that not only were developers always convinced that their intervention worked, but that generally all other parties did also. Put another way, there was no relationship between the eventual result of the impact evaluation and the views of stakeholders on whether the intervention worked or not. People involved just cannot tell whether something works or not, and therefore simply asking them if it works is no kind of evaluation at all. All such 'happy sheet' evaluations' should be ignored in future, and funders should cease using public and charitable money to fund them. The results can be very misleading, making them unethical in nature and harmful to the life chances of those that the work is intended to help.

A way forward for summarising and reporting evidence

The way in which the evidence from each trial is assessed in this paper is not (yet) widely used, although sensitivity analyses are becoming more common, and the star rating approach has been adapted with acknowledgement, and is now routinely used by the Educational Endowment Foundation (EEF) in England. This approach or something like it should be used. It is crucial to consider the design of any study in relation to its research question, and a robust evaluation requires something as powerful as a randomised controlled trial (or regression discontinuity design). It is also crucial to consider the scale, the method of allocation to groups and its success, the level of missing data (which must be reported scrupulously), the measurement quality, and other threats such as teaching to the test. And all of this must be clearly and carefully reported. If the report of an evaluation fails to explain any of these aspects then it is likely that the results are not trustworthy, and the evaluation has failed the test of trustworthiness. The NNTD is useful here in summarising two of these aspects of trustworthiness along with the 'effect' size.

There is certainly no role for significance testing or its hidden forms such as confidence intervals or multi-level modelling – now banned from use by many journals and areas of research. These techniques require complete randomisation of cases as a 'mathematical necessity' (Berk and Freedman), but they are routinely misused by unthinking researchers, in a way that is 'corrupt' (Starbuck 2016), and increasingly unethical (Gorard 2016). Estimating the p-value for any kind of non-random sample is pointless (Filho et al. 2013). The answer does not and cannot mean anything (Glass 2014). Even when used as intended these techniques do not provide the answer that most users want and imagine them to provide. No one wants to know the probabilistic answer that significance tests actually provide (Falk and Greenbaum 1995). And even if they worked as intended they would address none of the issues above – such as missing data, measurement quality and so on.

Why evaluation matters?

Evaluation is at the heart of any public policy where financial investments, such as the pupil premium, are set to achieve certain targets. Financial investment does not work unless there is a variety of useful choices for schools on how to use the funding to best effect. It would be unethical to continue to use a scheme that has been shown to be harmful or even just ineffective. It would be unwise to rely on a scheme that has not been repeatedly and robustly evaluated. Using professional judgment according to context and selecting an evidence-based approach from an array of possibilities is the ethical way forward. It is crucial that weak evaluations, such as those run by the developers of most interventions, or based on weak designs or high attrition rates, are not used to inform practice. Teachers and policy-makers need help to understand and appreciate the difference between weak and robust evidence of effectiveness.

References

Baenen, N., Bernhole, A. Dulaney, C. and Banks, K. (1997) Reading Recovery: Long-term progress after three cohorts, *Journal of Education for Student s Placed at Risk*, 2, 2, 161

BELL (2001) *BELL Accelerated Learning Summer Program 2001 evaluation report*, Dorchester, MA

BELL (2002) *BELL Accelerated Learning Summer Program 2002 national evaluation report*, Dorchester, MA

BELL (2003) *BELL Accelerated Learning Summer Program: 2003 program outcomes*, Dorchester, MA

Berk, R. and Freedman, D. (2001) *Statistical assumptions as empirical commitments*, http://www.stat.berkeley.edu/~census/berk2.pdf, accessed 030714

Borman, G., Benson, J. and Overman, L. (2009) A randomised field trial of the Fast ForWord Language computer-based training program, *Educational Evaluation and Policy Analysis*, 31, 82-106

Brooks, G. (2007) *What works for pupils with literacy difficulties? The effectiveness of intervention schemes*, London: DCSF Publications

Brooks, G., Harman, J. and Harman, M. (2003) *Catching Up at Key Stage 3: an evaluation of the Ruth Miskin [RML2] pilot project 2002/2003*, A report to the Department for Education and Skills, Sheffield: University of Sheffield

Bullock, J. (2005) *Effects of the Accelerated Reader on reading performance of third, fourth, and fifth-grade students in one western Oregon elementary school*, University of Oregon; 0171 Advisor: Gerald Tindal. DAI, 66 (07A), 56-2529

Chaplin, D. and Capizzano, J. (2006) *Impacts of a summer learning program: a random assignment study of Building Education Leaders for Life (BELL)*, Washington, DC: The Urban Institute, http://www.urban.org/UploadedPDF/411350_bell_impacts.pdf

Clark, C. (2013) *Accelerated Reader and young people's reading. Findings from the National Literacy Trust's 2012 annual literacy survey on reading enjoyment, reading behaviour outside class and reading attitudes*, London: National Literacy Trust, http://www.literacytrust.org.uk/assets/0001/9353/AR_and_young_people_s_reading.pdf

Clay, M. (1991) Becoming literate: The construction of inner control. Auckland: Heinemann

Coles, J. (2012) *An evaluation of the teaching assistant led Switch-on literacy intervention*, Unpublished MA thesis, University of London Institute of Education

Colom, R., Moriyón, F., Magro, C. and Morilla, E. (2014) The Long-term Impact of Philosophy for Children: A Longitudinal Study (Preliminary Results).*Analytic Teaching and Philosophical Praxis*, *35*, 1

Cooper, H., Charlton, K., Valentine, J. and Muhlenbruck, L. (2000) *Monographs of the Society for Research into Child Development*, 65, 1

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., et al. (2007) *Effectiveness of reading and mathematics software products: findings from the first pupil cohort*, (Publication No. 2007-4005), Washington, DC: U.S, Department of Education, Institute of Education Sciences, available from http://ies.ed.gov/ncee/pdf/20074005.pdf

Faggella-Luby, M., and Wardwell, M. (2011). RTI in a middle school: Findings and practical implications of a Tier 2 reading comprehension study. Learning Disabil-ity Quarterly, 34(1), 35–49.

Fair, F., Haas, L., Gardosik, C., Johnson, D. , Price, D. and Leipnik, O. (2015) Socrates in the schools from Scotland to Texas: Replicating a study on the effects of a Philosophy for Children program. *Journal of Philosophy in Schools*, 2(1)

Falk, R. and Greenbaum, C. (1995) Significance tests die hard: the amazing persistence of a probabilistic misconception, *Theory and Psychology*, 5, 75-98

Filho, D., Paranhos, R., da Rocha, E. , Batista, M., da Silva, J., Santos, M. and Marino, J. (2013) *When is statistical significance not significant?*, http://www.scielo.br/pdf/bpsr/v7n1/02.pdf

Fountas, I., and Pinnell, G. (1996) Guided reading: Good first teaching for all children. Portsmouth, NH: Heinenmann.

Fountas, I., and Pinnell, G. (2006) Teaching for comprehending and fluency: Thinking, talking and writing about reading, K-8. Portsmouth, NH: Heinenmann.

Galton, M., Gray, J., and Ruddock, J. (1999). T*he impact of school transitions and transfers on pupil progress and attainment*, DfEE Research Report No. 131. Norwich: HM's Stationery Office

Galton, M., Morrison, I. & Pell, T. (2000). 'Transfer and transition in English schools: reviewing the evidence' (Special Edition: School Transitions and Transfers), *International Journal of Educational Research*, 33, 4, 341-363.

Glass, G. (2014) Random selection, random assignment and Sir Ronald Fisher, *Psychology of Education Review*, 38, 1, 12-13

Goldacre, B. (2012) *Bad Pharma*, London: HarperCollins

Goolsbee, A. and Guryan, J. (2005) The impact of internet subsidies for public schools, *Review of Economics and Statistics*, 88, 2, 36-347

Gorard, S. (2013) *Research Design: Robust approaches for the social sciences*, London: SAGE

Gorard, S. (2014) A proposal for judging the trustworthiness of research findings, *Radical Statistics*, 110, 47-60, http://www.radstats.org.uk/no110/Gorard110.pdf

Gorard, S. (2016) Damaging real lives through obstinacy: re-emphasising why significance testing is wrong, *Sociological Research On-line*, 21, 1, http://www.socresonline.org.uk/21/1/2.html

Gorard, S. and Gorard, J. (2015) What to do instead of significance testing? Calculating the 'number of counterfactual cases needed to disturb a finding', *International Journal of Social Research Methodology*, 19, 4, 481-489

Gorard, S., See, BH and Davies, P. (2011) *Do attitudes and aspirations matter in education?: A review of the research evidence*, Saarbrucken: Lambert Academic Publishing

Gorard, S., Siddiqui, N. and See, BH (2015a) An evaluation of the 'Switch-on reading' literacy catch-up programme, *British Educational Research Journal*, 41, 4, 596-612

Gorard, S., Siddiqui, N. and See, BH (2015b) How effective is a summer school for catch-up attainment in English and maths?, *International Journal of Educational Research*, http://www.sciencedirect.com/science/article/pii/S0883035515301932

Gorard, S., Siddiqui, N. and See, BH (2016a) Can 'Philosophy for Children' improve primary school attainment, *Journal of Philosophy of Education*, (submitted)

Gorard, S., Siddiqui, N. and See, BH (2016b) An evaluation of Fresh Start as a catch-up intervention: And whether teachers can conduct trials, *Educational Studies*, 42, 1, 98-113

Gov.UK (2012) £10 million to boost literacy for year sevens, https://www.gov.uk/government/news/10-million-to-boost-literacy-for-year-sevens, Accessed 18/3/14

Graham, C. and Hill, M. (2003) *Negotiating the transition to secondary school*: SCRE Spotlight, Scottish Council of Research in Education, Edinburgh, http://files.eric.ed.gov/fulltext/ED482301.pdf

Graves, A., Brandon, R., Duesbery, L., McIntosh, A., and Pyle, N. (2011). The effects of Tier 2 literacy instruction in sixth grade: Toward the development of a re-sponse-to-intervention model in middle school. Learning Disability Quarterly, 34(1), 73–86. (full paper not available, analysis based on abstracts).

Harvard Family Research Project (2006) *Evaluation of the BELL (Building Educated Leaders for Life) Accelerated Learning Summer Program*, http://www.hfrp.org/out-of-school-time/ost-database-bibliography/database/bell-accelerated-learning-summer-program/evaluation-1-2002-national-evaluation-report

Institute of Education Sciences (IES) (2008) *What Works Clearing House Intervention Report: Accelerated Reader*, US Department of Education

Khan, M. and Gorard, S. (2012) A randomised controlled trial of the use of a piece of commercial software for the acquisition of reading skills, *Educational Review*, 64, 1, 21-36

Krashen, S. (2007) Accelerated Reader: Once again, evidence still lacking. *Knowledge Quest 36 September/October*, Available at: http://www.ala.org/aasl/aaslpubsandjournals/knowledgequest/kqwebarchives/v36/361/361krashen

Lanes, D., Perkins, D., Whatmuff, T., Tarokh, H. and Vincent, R. (2005) *A survey of Leicester City Schools using the RML1 and RML2 literacy programme*, Leicester: Leicester City LEA (mimeograph)

Leroux, A., Vaughn, S., Roberts, G., and Fletcher, J. (2011). Findings from a three-year treatment within a response to intervention framework for students in grades 6 with reading difficulties. Paper Presented at the Society for Research on Educa-tional Effectiveness Conference http://www.eric.ed.gov/PDFS/ED518866.pdf

Lipman, M., Sharp, A. and Oscanyon, F. (1980) *Philosophy in the classroom: Appendix B*, Philadelphia: Temple University Press

Mathis, D. (1996). *The Effect of the Accelerated Reader Program on Reading Comprehension*, US Department of Education, http://files.eric.ed.gov/fulltext/ED398555.pdf

Matsudaira, J. (2008) Mandatory summer school and student achievement, *Journal of Econometrics*, 142, 2, 829-850

May, H., Gray, A., Gillespie, J., Sirinides, P., Sam, C., Goldsworthy, H., Armijo, M. and Tognatta, N. (2013) *Evaluation of the i3 scale-up of Reading Recovery*, University of Delaware

Mercer, N., Wegerif R. and Dawes, L. (1999) Children's talk and the development of reasoning in the classroom, *British Educational Research Journal*, 25, 1, 95-111

Nichols, J. (2013) *Accelerated Reader and its effect on fifth-grade students' reading comprehension (*Doctoral dissertation, Liberty University)

OFSTED (2010) *Reading by six: How the best schools do it*, London: OFSTED

Pinnell, G., DeFord, D. and Lyons, C. (1988) *Reading Recovery: Early intervention for at-risk first graders*, Educational Research Service Monograph, Arlington, VA: Educational Research Service.

Pinnell, G., Lyons, C., DeFord, D., Bryk, A. and Seltzer, M. (1994) Comparing instructional models for the literacy education of high risk first graders, *Reading Research Quarterly*, 29, 1, 8-39

Reyes, O., Gillock, K. Kobus, K. and Sanchez, B. (2000) A longitudinal examination of the transition into senior high school for adolescents from urban, low-income status, and predominantly minority backgrounds, *American Journal of Community Psychology*, 28, 4, 519–44

Ross, S., Nunnery, J. and Goldfeder, E. (2004) A randomized experiment on the effects of Accelerated Reader/Reading Renaissance in an urban school district: Preliminary evaluation report. Memphis, TN: The University of Memphis, *Centre for Research in Educational Policy*

Rouse, C., and Krueger, A. (2004) Putting computerized instruction to the test: A randomised evaluation of a "scientifically-based" reading program, *Economics of Education Review*, 23, pp. 323–338

Sainsbury, M., Whetton, C., Keith, M. and Schagen, I. (1998) Fallback in attainment on transfer at age 11: evidence from the Summer Literacy Schools evaluation, *Educational Research*, 40, 1, 73-81

Schacter, J. and Jo, B. (2005) Learning when school is not in session: a reading summer day-camp intervention to improve the achievement of exiting First-Grade students who are economically disadvantaged, *Journal of Research in Reading*, 28, 2, 158–169

Schwartz, R. (2005) Literacy learning of at-risk first-grade students in the Reading Recovery early intervention, *Journal of Educational Psychology*, 97, N2, 257-26

Scott, L. (1999) *The Accelerated Reader program, reading achievement, and attitudes of students with learning disabilities,* Unpublished doctoral dissertation, Georgia State University, Atlanta (ERIC Document Reproduction Service No. ED 434431)

See, BH and Gorard, S. (2014) Improving literacy in the transition period: a review of the existing evidence on what works, *British Journal of Education, Society and Behavioural Sciences*, 4, 6, 739-754, http://www.sciencedomain.org/issue.php?iid=431andid=21

See, BH, Gorard, S. and Siddiqui, N. (2015) Best practice in conducting RCTs: Lessons learnt from an independent evaluation of the Response-to-Intervention programme, *Studies in Educational Evaluation*, 47, 83-92, http://www.sciencedirect.com/science/article/pii/S0191491X15000619

Shannon, L., Styers, M., Wilkerson, S., and Peery, E. (2015) Computer-assisted learning in elementary reading: A randomized control trial, *Computers in the Schools*, 32, 1, 20–34, doi: 10.1080/07380569.2014.969159

Siddiqui, N., Gorard, S. and See, BH (2014) Is a summer school programme a promising intervention in preparation for transition from primary to secondary school?, *International Education Studies*, 7, 7, 125-135, http://www.ccsenet.org/journal/index.php/ies/article/view/38214/21358

Siddiqui, N., Gorard, S. and See, BH (2015) Accelerated Reader as a literacy catch-up intervention during the primary to secondary school transition phase, *Educational*

Review, http://www.tandfonline.com/doi/full/10.1080/00131911.2015.1067883#.Vbs-W0_bKUk

Starbuck, W. (2016) 60th Anniversary Essay: How Journals Could Improve Research Practices in Social Science, *Administrative Science Quarterly*, 61, 2, 165–183

Tanner, E., Brown, A., Day, N., Kotecha, M., Low, N., Morrell, G., Turczuk, O., Brown, V., Collingwood, A., Chowdry, H., Greaves, E., Harrison, C., Johnson, G. and Purdon, S. (2011) *Evaluation of Every Child a Reader*, London: NatCen

Terzian, M. and Moore, K. (2009) *What works for summer learning programs for low-income children and youth?*, Washington: Child Trends, http://www.wallacefoundation.org/knowledge-center/summer-and-extended-learning-time/summer-learning/Documents/Effective-and-Promising-Summer-Learning-Programs-Fact-Sheet.pdf

Topping, K. (2014) *What kids are reading: The book reading habits of students in British Schools 2014*: An Independent Study, Renaissance Learning Inc.: United Kingdom

Topping, K. , and Trickey, S. (2007) Collaborative philosophical inquiry for schoolchildren: Cognitive gains at 2-year follow-up, *British Journal of Educational Psychology*, 77(4), 787-796

Trickey, S. and Topping, K. (2004) Philosophy for children: a systematic review, *Research Papers in Education*, 19, 3, 365-380

Vaughn, S., and Fletcher, J. (2012) Response to intervention with secondary school students with reading difficulties, *Journal of Learning Disabilities*, 4, 3, 244–256

West ,P., Sweeting, H. & Young R. (2010) Transition matters: pupils' experiences of the primary–secondary school transition in the West of Scotland and consequences for well- being and attainment, *Research Papers in Education*, 25:1, 21-50

What Works Clearinghouse (2013) Reading Recovery, http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=420